FMRI Data Analysis: Linear Models and Statistical Inference

Robert W Cox, PhD
Scientific and Statistical Computing Core
National Institute of Mental Health
National Institutes of Health
Department of Health and Human Services
Bethesda, MD USA

Conclusion First (so you don't have to read so much)
There are many ways to analyze FMRI datasets. FMRI-based investigators need to be aware of the different techniques, their underlying assumptions about the FMRI signal and noise, their strengths and limitations, and their applicability to any given experimental situation.

In other words: ***understand what you are doing.*** *Or you will do something stupid.*

Summary Second (if you are still reading)
The linear models used in the vast majority of FMRI-based papers are based on two assumptions: [*i*] *multiple repetitions of the same stimulus will result in the same response in the MRI signal ("shift invariance"), and* [*ii*] *when responses from multiple stimuli overlap in time, the signal changes add ("linearity"). Statistical inference from these models is based on the assumption that the noise is additive, Gaussian, and independent of the BOLD signal. The principal differences between various linear analysis methods lie in* [*a*] *the modeling of the temporal shape of the BOLD response, and* [*b*] *the assumptions about the spatial and temporal correlation of the noise.*

Signal Modeling Principles in FMRI Data Analysis

A *signal* is a measurable response, often a response to a stimulus; *noise* is the component of measurement that interferes with detection of the signal. Statistical decision theory is a branch of mathematics/signal processing/statistics that deals in development of methods to distinguish noise-only measurements from signal+noise measurements. This theory requires understanding the relationship between the stimulus and the signal, and requires characterizing the noise statistically.

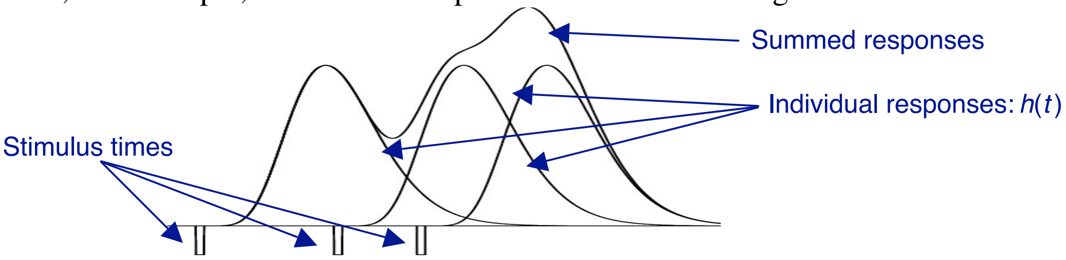
FMRI datasets are particularly challenging. The stimulus-signal relationship and the noise statistics are both poorly characterized and are both still subjects of research. The result is that there is no "best" way to analyze FMRI time series data: there are only "reasonable" analysis methods. It is often appropriate to analyze a dataset in more than one way and then compare the results to see if the neuroscience conclusions are affected. It is a measure of the robustness of FMRI datasets that when such comparisons are made, the partisans for various methods are usually reduced to arguing about very minor changes in the activation "blobs".

To deal with data systematically, we must make some assumptions about the signal and the noise. These assumptions will be wrong (overly simple), which means that it is important to understand the model underlying any given analysis, and perhaps also to try more than one analysis method to see if the results vary significantly. Different kinds of experiments will require different kinds of analyses, since the signal models and questions asked about the signal will vary.

One fundamental problem with FMRI data analysis is that we don't have enough data. This sounds crazy: it is routine now to gather 1+ Gigabytes of data per subject. But most of this vast pile of bits is not relevant to neural activity (the BOLD component of the FMRI signal is weak), and we must make many decisions to make a brain map. Typically, there are 10,000-100,000 EPI voxels inside the brain, and at least one decision is made per voxel (e.g., "*is it active*?"). If the chance of an error is 1% per voxel, then we'd expect 100-1000 errors in every brain map. This may be as big as the number of truly active voxels in the brain; such results would be garbage. Adapting to this "curse of multiple comparisons" is a major issue in FMRI data analyses.

Temporal Models: Linear Convolution

A *linear* model is built on the assumption that the signals from separate stimuli just add up. A *convolution* model assumes the separate signals from identical stimuli are the same in shape, amplitude, and in time delay from the stimulus times. Linear convolution models are the most common in FMRI; for example, an idealized response to three stimuli might look like:



In the absence of noise, the "summed responses" curve, plus a *baseline,* would comprise our measurements.

A large number of linear convolution measurement models can be written down, each one customized to the particular experimental application. For example, if we assume that the stimuli occur on the *TR* time grid, we could write

$$\underbrace{Z(t)}_{\substack{\text{voxel data}\\ \text{at time } t}} = \underbrace{\beta_0+\beta_1\cdot t}_{\substack{\text{baseline model}\\ \text{(linear in time)}}} + \sum_{m=0}^{p} \underbrace{f(t-m\cdot TR)}_{\substack{\text{stimulus}\\ \text{function (0 or 1)}\\ m\cdot TR \text{ in the past}}} \cdot \underbrace{h(m\cdot TR)}_{\substack{\text{hemodynamic}\\ \text{response func}\\ m\cdot TR \text{ after stim}}} + \underbrace{\varepsilon(t)}_{\text{noise}}$$

Here, the stimuli are modeled as off or on (0 or 1); the response at time t to an individual stimulus that occurs at time 0 is denoted by the hemodynamic response function (HRF) $h(t)$ (i.e., a stimulus at time τ causes a response $h(t—\tau)$ at time t); and the baseline is modeled as a linear function of time. The sum extends $p\cdot TR$ into the past; the length of this interval is chosen to match the expected duration of the MRI signal response after the stimulus.

If we assume that the stimulus times are not bound to the *TR* grid, then a related model is

$$Z(t)=\beta_0+\beta_1\cdot t+\sum_{s=1}^{N_s} h(t-\tau_s)+\varepsilon(t)$$

where there are N_s stimuli that occur at times τ_1, τ_2, …. A similar model can be used for a stimulus that has two phases, which don't always occur at the same temporal offset (e.g., the first phase is a visual presentation and the second phase is the subject response):

$$Z(t)=\beta_0+\beta_1\cdot t+\sum_{s=1}^{N_s}\left[h_1\left(t-\tau_s\right)+h_2\left(t-(\tau_s+\delta_s)\right)\right]+\varepsilon(t)$$

where δ_s is the delay ("jitter") between the two phases for the s^{th} stimulus.

All of these models can be generalized to allow for multiple classes of stimuli. Each stimulus class q would get its own hemodynamic response function $h^{(q)}(t)$.

We have not yet specified precisely what we are trying to determine from the data $Z(t)$; something about $h(t)$ presumably, but what? There are two classes of hemodynamic models that are widely used in FMRI data analysis: fixed shape HRF with only the amplitude of response unknown, and parameterized ("variable shape") HRF with the shape and amplitude of response unknown. (N.B.: What we actually observe is derived from the hemodynamic impulse response function convolved with the neuronal response function, and that making inferences about one of these functions alone require making assumptions about the other function, or making some additional measurements.)

Fixed Shape HRF: In these models, we assume $h(t)=\alpha\cdot r(t)$, where α is unknown and $r(t)$ is some reference function we choose; Mark Cohen's function $r(t)=t^b e^{-t/c}$ for $t>0$ is a popular shape (e.g., b=8.6 and c=0.547; the time delay to the peak is $b\cdot c$ and the FWHM of the peak is approximately $2.4\cdot c\cdot\sqrt{b}$ for $b>1$). These models have the advantage of having fairly few parameters per voxel: one $\alpha^{(q)}$ for each stimulus class q, plus the baseline parameters (β_0 and β_1 in the models above). The b and c parameters are fixed in these types of models, and are often assumed to be the same for all subjects. A refinement is to separately estimate b and c for each individual using a simple motor or visual FMRI paradigm, prior to the more complicated experiment that you are undoubtedly contemplating.

The BOLD response to a brief stimulus (e.g., a 100 ms flash of light) typically lasts about 10-12 seconds, comprising a 2 s delay, 3-5 s rise and a 4-5 s fall. For long values of *TR* (3 s or more), using a fixed shape HRF makes a great deal of sense: there isn't enough temporal resolution to try to capture the shape.

Variable Shape HRF: In these models, more parameters are added to the unknown function $h(t)$ in order to let its shape vary. There are two principal motivations: first, to fit the data $Z(t)$ better in each voxel so that the statistical significance of activation is properly assessed; and second, to allow statistical inference on the shape of $h(t)$ itself (e.g., is the activation amplitude stronger from 4-8 s post-stimulus or from 8-12 s post-stimulus?).

For example, the widely-used standard SPM variable shape HRF model has the form $h(t)=\alpha_0\cdot r_{pk}(t)+\alpha_1\cdot r'_{pk}(t)-\alpha_2\cdot r_{pu}(t)-\alpha_3\cdot r'_{pu}(t)$, where each $r_{xx}(t)$ is of Cohen's form, with $r_{pk}(t)$ using (b,c) parameters that represent the BOLD peak and $r_{pu}(t)$ using (b,c) parameters representing the later BOLD post-undershoot. The inclusion of the derivatives $r'_{xx}(t)$ allows for small unknown time shifts, since $r_{xx}(t+s)\approx r_{xx}(t)+s\cdot r'_{xx}(t)$. (If there is more than one stimulus class, each class requires a separate set of four α parameters.)

More complicated models (e.g., polynomial, spline, or trigonometric series) can allow for more shape flexibility. However, it can become impossible to find activation (i.e., state confidently that $h(t)$ is nonzero) when there are too many parameters for the data, since a very high-dimensional model will fit a pure noise voxel almost as well as it fits a signal+noise case. Similar problems arise when the number of stimulus classes is increased; again, the number of parameters increases (a few α's per q), and it is easy to go too far. It is best to start with a simple analysis of FMRI time series data, see if the results make sense, then progress to the use of more complicated models to extract more information. In this way, the data analyst can get a feel for how many parameters can be estimated from the datasets. It is a common mistake to group the stimuli into too many classes, so that there are relatively few (under 20, say, in an event-related design) responses per class. FMRI datasets are not good enough to reliably assess differential activation between tasks when there are so few samples per task.

Inverse Models: Instead of solving for $h(t)$ in each voxel, one can assume a fixed $h(t)$ and then solve for the stimulus time series $f(t)$ that best fits the data in each voxel. This approach has the potential for finding neural activation patterns for complex continuous stimuli such as video or audio presentations. Such inverse models have not been widely applied, partly because they involve a large number of parameters per voxel (for fitting $f(t)$ to the data $Z(t)$).

Statistical Inference from Linear Models: Under the assumption that the noise is Gaussian and has a known temporal correlation structure, the statistics of linear models are fairly straightforward. The unknown parameters are estimated using a least-squares fitting criterion (e.g., minimize $E = \sum_t \left| Z(t) - \text{model}(t) \right|^2$). The magnitude of E is used to estimate the variance of the noise. From these estimates, the significance of linear combinations of the model parameters can be calculated directly using F- or t-statistics (F-tests for multiple combinations of parameters, t-tests for single combinations). For example, in the fixed shape model, where the only parameter of activation interest is α, the test gives the probability p that $\alpha=0$ given the data. If this p-value is sufficiently small, we declare this voxel to be "active" and colorize it somehow (usually, the color is based on the amplitude α, but is sometimes instead based on the F- or t-statistic). If we had two stimulus classes and so estimated $\alpha^{(1)}$ and $\alpha^{(2)}$ as the response magnitudes for each type of stimulus, we could determine the p-value for the null hypothesis $\alpha^{(1)}-\alpha^{(2)}=0$; we would presumably colorize voxels in which this p was small (indicating that $\alpha^{(1)}\neq\alpha^{(2)}$) *and* the p-values for $\alpha^{(1)}=0$ and/or $\alpha^{(2)}=0$ were also small—these would be locations where the brain responded to at least one of the types of stimuli *and* responded differently to the two different stimulus types. And so forth—this kind of test is sometimes called a *conjunction analysis*. The limits of this type of inference are your imagination. And the quality of the data.

Nonlinear Models? There is nothing wrong with using nonlinear models for FMRI time series; for example, one could directly solve for the (b,c) parameters in Cohen's model, in each voxel. The practical drawback to nonlinear models is the difficulty of solving the fitting equations for the parameters. Linear models have the strong advantage that the least-squares criterion leads to linear equations for the unknown parameters; efficient algorithms for solving such equations have been well-established since the 1960s. The same cannot be said for solving nonlinear optimization problems. Nevertheless, nonlinear models have some attractive features, such as providing the ability to impose constraints on the shape of the expected response.

Spatial Models

The most common form of FMRI data analysis is voxel-wise: each voxel time series $Z(t)$ is analyzed separately from all others. The attraction is that the full spatial resolution of the echo-planar images is kept. However, we probably wouldn't accept a brain activation map that consisted solely of randomly scattered "on" voxel with no clear spatial structure; instead, we'd go back to the data and try to figure out what went wrong. But if we aren't going to accept an arbitrary spatial map, then we can increase our statistical power by only looking for spatial activation patterns that are "reasonable". There are three commonly-used methods.

Smoothing: One of the simplest ways to produce "reasonable-looking" activation maps is to smooth the FMRI data spatially prior to the temporal analysis (or maybe after analysis). If a 10–15 mm FWHM Gaussian blur is used for this smoothing, for example, then FMRI results can be made to look much like PET results. The drawback to such simple smoothing is obvious: why bother to acquire high-resolution images if the first thing one does is to throw that resolution away? "Smart" smoothing is a variation that only does blurring within the gray matter (e.g., as detected from a T1-weighted volume with ≈1 mm resolution). This technique uses the high resolution of FMRI cleverly.

Clusters: A second way to produce "reasonable-looking" activation maps is via a dual-thresholding technique. After a voxel-wise time series analysis, voxels in which $h(t)$ is significantly different from 0 are selected; this first significance threshold is taken to be low, so that a fair number of false positives can be expected. The second thresholding step is to only accept contiguous clusters of voxels that passed the first step; the threshold here is the minimum allowable cluster size. This technique allows for the detection of relatively small amplitude activations, provided these activations cover a large region.

Regions of Interest (ROIs): A third method is to pre-select voxels that are to be averaged together; the selection is usually based on some anatomical criterion (e.g., the left hippocampus). This technique has the advantage that specifically targeted anatomical hypotheses can be addressed precisely, and that the regions can be tailored to each subject's anatomy. It has the disadvantage that intra-ROI differences can be lost (e.g., anterior vs. posterior hippocampus, if the ROI averages over the entire structure). It also has the disadvantage that manually selection of ROIs is a very time-consuming task. Even postdocs have been known to rebel at this. (At the NIH, we now have "postbacs" to do the work that postdocs won't do.)

Statistical Inference: The major point of using spatial models is that they reduce the multiple comparisons problem. In the case of ROI analyses, it is often the case that only 10-20 ROIs are used; the problem of dealing with 10,000-100,000 comparisons has been tossed away. In the case of smoothing, the elaborate analysis of "correlated random fields" has been developed to determine the analytical relationship between F- and t-statistic thresholds and p-values when the noise is strongly spatially correlated. In the case of clustering, the analytical relationship between cluster size threshold and p-value is unknown; as a result, direct numerical simulation (i.e., of how big a cluster might get if there is no signal, only noise) is usually used.

Noise Models and Regression Methods

Modeling the distribution of the noise $\varepsilon(t)$ is nearly as important as modeling the signal (but not as much fun). It is almost always assumed that the noise is Gaussian—normally distributed—mostly as a matter of analytical convenience and also as a practical starting point. The issue then becomes one of modeling the correlation structure of the noise.

For most forms of FMRI, the dominant source of "noise" in the time series is physiological fluctuations: the subject's heartbeat and breathing both contribute strongly to the variance in the data—usually 3-4 times as much as the intrinsic MRI measurement variance. The signals from such quasi-regular sources are correlated in time and space. For the most part, the spatial correlations are ignored in FMRI analyses; the temporal correlations are another story.

White Noise: The simplest analytical assumption is that the noise is uncorrelated in time and space—this is the meaning of the term "white". However, statistical inference about whether

$h(t)\neq 0$ will tend to be optimistic (i.e., the p-values will be too small) when the temporal correlations are strong.

Colored Noise: One approach to dealing with non-white noise is to attempt to reduce the degrees-of-freedom in the statistical estimates to allow for the fact that noise samples at successive times are not independent. Another approach is to modify the statistical estimation procedure for $h(t)$ to decorrelate ("prewhiten") the time series—this will preserve the degrees-of-freedom. Both methods require estimating the temporal correlation structure in some way.

Filtering: A related technique is to temporally filter the data. Removing high-frequency components makes sense, since we know the BOLD effect takes 8-10 s to play out, anything above (say) 0.1 Hz must be noise. Also, removing low-frequency drifts makes sense, since these are commonly seen in FMRI experiments (say, anything below 0.01 Hz, or some frequency well below the stimulation frequency band). After this is done, we've imposed our own temporal correlation structure on the data in addition to the physiological structure. It is plausible to say that the filter-induced correlations are larger than the physiology-induced correlations; we can then adjust the degrees-of-freedom downwards for the filter-induced correlation structure and let it go at that.

Spectral Resampling: Another way to estimate the significance of an estimated parameter is via data randomization. We want to determine if the value of the parameter that we estimated was likely to have arisen just from noise. To do this, we could simply generate many samples of noise-only simulations, then analyze this "data". The difficulty is generating realistic noise when one of the issues is that we are admitting ignorance about the structure and distribution of the noise. *Resampling* methods (e.g., randomization, bootstrap) deal with this by using the data itself as the source of the noise-only simulations. In the simplest case, the data time series can be scrambled (in time) and then re-analyzed. Scrambling ("randomization") will destroy the stimulus-response link (good); however, it will also destroy the temporal correlation of the noise (bad). A way around this latter difficulty is to scramble the data only after transforming it to a spectral domain (e.g., wavelets, or short-time Fourier transforms). Spectral transforms tend to decorrelate noise—the correlation is mostly expressed in the different magnitudes of the spectral coefficients. The coefficients are then scrambled appropriately, the inverse spectral transform is made, and *voila!*—a noise-only simulation with the (mostly) correct temporal correlation structure has been produced.

Direct Physiological "Noise" Reduction: It is certainly possible to monitor the heartbeat and breathing of the subject during the FMRI experiment (e.g., with EKG and a respiratory belt "). It is then possible to partially filter out the components of the FMRI time series that are correlated with the physiological reference data. Ideally, this should be done on the complex-valued MR data, since a significant part of the noise turns out to be in the phase of the data. Unfortunately, some manufacturers make it difficult or impossible to obtain the complex-valued data.

Regression Methods: As mentioned earlier, least-squares regression is by far the most common method used to fit models to data time series. This has two advantages: it is simple, and it is optimal when the noise is truly Gaussian. However, if the noise is not Gaussian, least-square regression can be severely influenced by a few "outliers"; in least squares regression, a datum that is 10 standard deviations away from the model counts 100 times as much in the calculation of the fitting error E as does a point that is 1 standard deviation away. With Gaussian noise, the probability that a point is 10 or more standard deviations out is about 10^{-23}, so we don't worry about it much. But such points do sometimes occur in FMRI time series, often due to scanner hardware glitches. It is often worth examining FMRI time series for outliers, since even a few can destroy an otherwise useful dataset.

One way to minimize such problems, while not explicitly going “outlier-hunting”, is to use a more robust regression method—one that is less sensitive to a few wild points. In other fields, a popular alternative to least-squares regression is L^1 fitting, where the criterion is to minimize $E = \sum_t |Z(t) - \text{model}(t)|$. The algorithms for this fitting are also well-established, provided that the model is linear in the unknown parameters. The main objection to using L^1 regression is that the statistics are less well understood; the parameter estimates are asymptotically Gaussian as the number of data points goes to infinity, but it is not clear how relevant that is to the cases that arise in FMRI data analysis. Nor is it clear how to deal with temporal correlations in the noise with L^1 regression, except perhaps by some resampling scheme.

Spatially Structured Noise: As alluded to earlier, if the noise is strongly correlated across voxels, this fact can be used (via the theory of correlated random fields) to reduce the effective number of comparisons needed to make an activation map. Another way to use the fact that the voxels are related is to try to estimate the temporal correlation structure in each voxel time series using not just that voxel’s data but also the data from neighboring voxels. In this methodology, spatial smoothing is being applied, not directly in the estimation of the response magnitude α, but rather in the estimation of the properties of the noise (i.e., any smoothing is applied only after the α’s are estimated and their effect is subtracted from the data, leaving only the *residuals* behind). The impetus for using such techniques comes from the fact that accurate estimation of temporal correlation structure is difficult since FMRI time series are usually fairly short (e.g., 100 points). Assuming that neighboring voxels have similar temporal correlations helps to overcome this problem.

Conclusion Again

There are many ways to analyze FMRI datasets.

- As seen above, even the linear shift-invariant model has many variations (and there are more than this outline has covered!), each of which would give different results—only slightly different, we would hope.

FMRI-based investigators need to be aware of the different techniques, their underlying assumptions about the FMRI signal and noise, their strengths and limitations, and their applicability to any given experimental situation.

- One might hope that FMRI data analysis would be a “one size fits all” situation, but that is not the case. Just as the experiment design must be tailored to explore your underlying hypotheses, so the data analysis must be tailored to answer the questions that you pose from the data that you have.
- People argue about the underlying assumptions that are explicit or implicit in the various methods that are outlined above, debating both the validity of the assumptions and their importance. To judge between these methods requires (at least) a conceptual appreciation of the techniques.

*In other words: **understand what you are doing.** Or you will do something stupid.*

- Linear analysis is a complex and powerful tool, and “with great power comes great responsibility.” In this case, the responsibility is to understand.